Towards a sparse, scalable, and stably positive definite (inverse) covariance estimator

Sang-Yun Oh

Department of Statistics and Applied Probability, University of California
Santa Barbara, California 93106-3110, U.S.A.
syoh@pstat.ucsb.edu

Bala Rajaratnam

Department of Statistics, Stanford University
390 Serra Mall, Stanford, California 94305-4065, U.S.A.
brajarat@stanford.edu.

Joong-Ho Won

Department of Statistics, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
wonj@stats.snu.ac.kr.

Abstract

High dimensional covariance estimation and graphical models is a contemporary topic in statistics and machine learning having widespread applications. The problem is notoriously difficult in high dimensions as the traditional estimate is not even positive definite. An important line of research in this regard is to shrink the extreme spectrum of the covariance matrix estimators. A separate line of research in the literature has considered sparse inverse covariance estimation which in turn gives rise to graphical models. In practice, however, a sparse covariance or inverse covariance matrix which is simultaneously well-conditioned and at the same time computationally tractable is desired. There has been little research at the confluence of these three topics. In this paper we consider imposing a condition number constraint to various types of losses used in covariance and inverse covariance matrix estimation. This extends the approach by Won, Lim, Kim, and Rajaratnam (2013) on multivariate Gaussian log likelihood. When the loss function can be decomposed as a sum of an orthogonally invariant function of the estimate and its inner product with a function of the sample covariance matrix, we show that a solution path algorithm can be derived, involving a series of ordinary differential equations. The path algorithm is attractive because it provides the entire family of estimates for all possible values of the condition number bound, at the same computational cost of a single estimate with a fixed upper bound. An important finding is that the proximal operator for the condition number constraint, which turns out to be very useful in regularizing loss functions that are not orthogonally invariant and may yield non-positive-definite estimates, can be efficiently computed by this path algorithm. As a concrete illustration of its practical importance, we develop an operator-splitting algorithm that imposes a guarantee of well-conditioning as well as positive definiteness to recently proposed convex pseudo-likelihood based graphical model selection methods (Zhang and Zou, 2014; Khare, Oh, and Rajaratnam, 2015).

1 Introduction

We consider the problem of estimating the covariance matrix or its inverse (precision matrix) from n independent copies of p-variate random vectors from some distribution. This estimation problem is becoming increasingly important in many statistical methods, from least squares reqression to graphical model selection. Applications include medical image analysis, genomics, and financial engineering, to name a few. In some applications (e.g., portfolio optimization, Gauss mixture clustering) overall risk properties of the covariance estimator are important; in others (e.g., graphical model selection), the sparsity pattern of the inverse covariance matrix is of critical interest. In any situation, the estimator should be symmetric, positive definite to be a valid (inverse) covariance matrix. It is also desirable that the ratios between the eigenvalues of the estimator are not too extremal, in order to reflect that the population covariance matrix describes a proper, non-degenerate p-dimensional distribution. In this paper, we call matrices that satisfy both conditions to be *stably* positive definite.

Unfortunately, however, many estimators of covariance or inverse covariance matrix are not positive definite, let alone stably positive definite. It is well known that the sample covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T, \tag{1}$$

where X_i is the ith copy of the random vector, is merely positive *semidefinite* when $n < p$. Some high-dimensional covariance matrix estimators based on structural sparsity assumptions may fail to be positive definite (Fan, Liao, and Mincheva, 2013); high-dimensional sparse inverse covariance matrix estimators based on maximum pseudo-likelihood principle (Meinshausen and Bühlmann, 2006; Peng, Wang, Zhou, and Zhu, 2009; Zhao, Rocha, and Yu, 2009; Khare, Oh, and Rajaratnam, 2015; Zhang and Zou, 2014) may have negative eigenvalues, sometimes not even symmetric.

The main subject of study in this paper is the set of positive definite matrices with bounded condition numbers. The condition number of a positive definite matrix quantifies its degree of invertiblity, and is defined as the ratio of the largest to smallest eigenvalues of the matrix. Thus

the set of interest can be formally written, for an upper bound κ,

$$\begin{aligned}\mathcal{C}_\kappa &= \{\Omega : \Omega \succ 0,\ \lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \leq \kappa\} \\ &= \{\Omega : \exists u > 0, uI \preceq \Omega \preceq \kappa u I\},\end{aligned}$$

where $A \succ 0$ (*resp.* $A \succeq 0$) denotes that matrix A is positive definite (*resp.* positive semidefinite), $A \preceq B$ means that $B - A \succeq 0$, and $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ refers to the maximum and minimum eigenvalues of A; the identity matrix is denoted by I. One should note that the set $\mathcal{C}_\kappa$ properly encodes the notion of stable positive definiteness. If the (inverse) covariance matrix can be estimated constrained on $\mathcal{C}_\kappa$, then the estimator possesses the desired properties mentioned in the previous paragraph. Because $\Omega \in \mathcal{C}_\kappa$ implies that Ω^{-1} exists and $\Omega^{-1} \in \mathcal{C}_\kappa$, we do not distinguish estimation of the covariance matrix and estimation of the inverse covariance matrix too much; for the reason that will become apparent in the sequel, we use Ω to denote the inverse covariance matrix. Won, Lim, Kim, and Rajaratnam (2013) studied the set $\mathcal{C}_\kappa$ as a means to regularize high-dimensional Gaussian maximum likelihood covariance estimators. Their motivation is to impose numerical stability for inversion of the estimates, for instance to use with Markowitz-type portfolio optimization problems. In this paper, we see this idea can be extended to a much general class of loss functions.

Now consider the estimation problem of the form

$$\begin{array}{ll}\text{minimize} & L(\Omega) - \mathbf{Tr}(\Omega f(S)) \\ \text{subject to} & \Omega \in \mathcal{C}_\kappa,\end{array} \tag{2}$$

where $L(\Omega)$ is convex; S is the sample covariance matrix (1); and f is a function that maps a symmetric matrix to a symmetric matrix of the same dimension. Problem (2) includes many interesting cases:

1. Gaussian log likelihood: $L(\Omega) = -\log\det\Omega$, $f(S) = -S$.

2. Gaussian log likelihood with a-pair-of-nuclear-norms regularization (Chi and Lange, 2014): $L(\Omega) = -\log\det\Omega + \eta(\alpha\|\Omega\|_* + (1-\alpha)\|\Omega^{-1}\|_*)$, $f(S) = -S$.

3. Quadratic loss: $L(\Omega) = (1/2)\|\Omega\|_F^2$, $f(S) = S$.

4. CONCORD loss (Khare et al., 2015): $L(\Omega) = -\log\det\Omega_D + (1/2)\mathbf{Tr}(\Omega S\Omega)$, $f(S) = 0$, where $\Omega_D = \mathrm{diag}(\Omega_{11}, \ldots, \Omega_{pp})$.

5. D-trace loss (Zhang and Zou, 2014): $L(\Omega) = (1/2)\mathbf{Tr}(\Omega S\Omega)$, $f(S) = I$.

Hence a characterization of the solution to (2) is of an utter interest. Cases 1 – 3 are distinguished from the rest because in these cases $L(\Omega)$ is orthogonally invariant, i.e., $L(Q^T\Omega Q) = L(\Omega)$ for any Q such that $Q^TQ = I$, with an additional condition that $L(D) = \sum_{i=1}^p l_i(d_i)$, l_i being closed convex, if $D = \mathrm{diag}(d_1, \ldots, d_p)$. For instance,

$$l_i(\lambda) = \begin{cases} -\log\lambda, & \text{case 1,} \\ -\log\lambda + \eta(\alpha\lambda + (1-\alpha)\lambda^{-1}), & \text{case 2,} \\ (1/2)\lambda^2, & \text{case 3.} \end{cases}$$

In such cases, we can provide a complete characterization of the solution path of (2) as the parameter κ varies from unity to infinity. Furthermore, we show that for many interesting cases, the entire solution path can be computed at the same cost (namely, in $O(p)$ operations) as that of finding the solution for a fixed κ. Thus the characterization of the solution path provides a huge computational advantage in solving (2) efficiently.

Cases 4 and 5 are pseudo-likelihood losses that arise in high-dimensional graphical model selection. Orthogonal variance of $L(\Omega)$ in these cases prevents a direct application of the method mentioned in the previous paragraph. Nevertheless we can show that problem (2) with these losses can be efficiently solved by a scalable, Dykstra's alternating projection-type operator splitting method (Lange, 2013), resulting in a sparse, stably positive definite covariance selection. This is because the orthogonal projection of a symmetric matrix to set $\mathcal{C}_\kappa$ has an almost closed form representation, a result that follows from Section 2. In this sense, case 3 bridges cases 1 and 2 with cases 4 and 5.

The rest of the paper is organized as follows. In Section 2, we characterize the solution path for the orthogonally invariant cases as soultions of ordinary differential equations with respect

to κ, introduce an efficient method to solve (2) for all values of κ based on this observation. Explicit solutions to some cases introduced in this section are also provided. In Section 3 we develop an alternating projection algorithm that solves the orthogonally variant cases scalably, and demonstrate that the algorithm provides stably positive semidefinite solutions to graphical model selection problems, without loosing the desired sparsity. Section 4 concludes this paper. Some proofs of the results in the paper are given in the Appendix.

2 Solution path for orthogonally invariant $L(\Omega)$

We begin with the characterization of the solution to (2) for a fixed κ.

Theorem 1. *Suppose the spectral decomposition of $f(S)$ is given by VDV^T, $V^TV = VV^T = I$, $D = \mathrm{diag}(d_1, \ldots, d_p)$, $d_1 \geq \cdots \geq d_p$. Then, $\Omega^\star = V\Lambda^\star V^T$ minimizes (2), where $\Lambda^\star = \mathrm{diag}(\lambda_1^\star, \ldots, \lambda_p^\star)$ with*

$$\lambda_i^\star = \max(u^\star, \min(\tilde{\lambda}_i, \kappa u^\star)). \tag{3}$$

The $\tilde{\lambda}_i$ is the minimizer of $l_i(\lambda) - d_i\lambda$ in $\lambda \geq 0$. Let $u_{\alpha,\beta} = \mathrm{argmin}_u\, l_{\alpha,\beta}(u)$ where

$$l_{\alpha,\beta}(u) = \sum_{i=1}^{\alpha} l_{p-i+1}(u) - u\sum_{i=1}^{\alpha} d_{p-i+1} + \sum_{i=\beta}^{p} l_{p-i+1}(\kappa u) - \kappa u\sum_{i=\beta}^{p} d_{p-i+1}$$

for $\alpha \in \{1, \ldots, p-1\}$ and $\beta \in \{2, \ldots, p\}$. Then $u^\star$ can be chosen to equal to $u_{\alpha,\beta}$ for (α, β) satisfying the relation

$$(u_{\alpha,\beta}, v_{\alpha,\beta}) \in R_{\alpha,\beta} = \{(u, v) : \tilde{\lambda}_{p-\alpha+1} < u \leq \tilde{\lambda}_{p-\alpha}, \tilde{\lambda}_{p-\beta+2} \leq v < \tilde{\lambda}_{p-\beta+1}\}, \quad v_{\alpha,\beta} = \kappa u_{\alpha,\beta}.$$

Finding the pair (α, β) takes $O(p)$ time.

The proof is given in Appendix 1.

Remark 1. *This theorem subsumes Won et al. (2013, Theorem 1) that corresponds to case 1, and allows $f(S)$ to be indefinite or singular, i.e., $d_i \leq 0$ for some i. Thus $\tilde{\lambda}_i = \infty$ or $\tilde{\lambda}_i = -\infty$ is*

allowed.

Remark 2. *An insepection of the proof reveals that the problem reduces to determine*

$$u^\star = \operatorname*{argmin}_{u>0} \sum_{i=1}^{p} l_i(\lambda_i^*(u)) - d_i\lambda_i^*(u),$$

where $\lambda_i^*(u) = \max(u, \min(\tilde{\lambda}_i, \kappa u))$*, i.e. a univariate minimization problem. Thus standard univariate optimization methods, e.g., bisection or golden search, can also be employed to find* $u^\star$*, subject to a tolerance level. The theorem says that it can be found* exactly *within* $O(p)$ *operations.*

If l_is are continuously differentiable, the $u_{\alpha,\beta}$ in Theorem 1 can be found by solving the equation

$$\sum_{i=1}^{\alpha} l'_{p-i+1}(u) + \kappa \sum_{i=\beta}^{p} l'_{p-i+1}(\kappa u) = \sum_{i=1}^{\alpha} d_{p-i+1} + \kappa \sum_{i=\beta}^{p} d_{p-i+1}. \tag{4}$$

Then the implicit function theorem states that $u_{\alpha,\beta} = u_{\alpha,\beta}(\kappa)$ is a continous function of κ. Thus if the optimal $u^\star$ in (3) satisifies $u^\star(\kappa) = u_{\alpha,\beta}(\kappa)$ so that $u_{\alpha,\beta}(\kappa), v_{\alpha,\beta} \in \mathbf{int}R_{\alpha,\beta}$ for some α, β, where $\mathbf{int}A$ denotes the interior of a set A, then a small change in κ will not change α or β, i.e., $u^\star(\kappa + \Delta\kappa) = u_{\alpha,\beta}(\kappa + \Delta\kappa)$ and $(u_{\alpha,\beta}(\kappa + \Delta\kappa), v_{\alpha,\beta}(\kappa + \Delta\kappa)) \in \mathbf{int}R_{\alpha,\beta}$ for sufficiently small $\Delta\kappa$. Thus the local solution path within $R_{\alpha,\beta}$ can be traced by solving (4) for continuously varying κ subject to the condition $u^\star(\kappa) \in \mathbf{int}R_{\alpha,\beta}$. If we further assume that l_is are twice differentiable, this local path can be completely characterized by an ordinary differential equation: it is straighforward to derive

$$\frac{du_{\alpha,\beta}}{d\kappa} = \frac{\sum_{i=\beta}^{p} d_{p-i+1} - \sum_{i=\beta}^{p} l'_{p-i+1}(\kappa u) - \kappa u \sum_{i=\beta}^{p} l''_{p-i+1}(\kappa u)}{\sum_{i=1}^{\alpha} l''_{p-i+1}(u) + \kappa^2 \sum_{i=\beta}^{p} l''_{p-i+1}(\kappa u)}, \tag{5}$$

from which the curve $(u^\star(\kappa), v^\star(\kappa))$ within $R_{\alpha,\beta}$ can be determined.

Example 1. *For case 1, we have*

$$\frac{du_{\alpha,\beta}}{d\kappa} = -\frac{(\alpha + p - \beta + 1)\sum_{i=\beta}^{p} s_i}{(\sum_{i=1}^{\alpha} s_i + \kappa \sum_{i=\beta}^{p} s_i)^2}, \tag{6}$$

where s_i is the ith largest eigenvalue of S. In this case (4) *has an explicit solution*

$$u_{\alpha,\beta}(\kappa) = \frac{\alpha + p - \beta + 1}{\sum_{i=1}^{\alpha} s_i + \kappa \sum_{i=\beta}^{p} s_i},$$

which satisfies (6)*. Furthermore, because*

$$\frac{dv_{\alpha,\beta}}{d\kappa} = \frac{(\alpha + p - \beta + 1)\sum_{i=1}^{\alpha} l_i}{(\sum_{i=1}^{\alpha} s_i + \kappa \sum_{i=\beta}^{p} s_i)^2},$$

it follows that

$$\frac{dv_{\alpha,\beta}}{du_{\alpha,\beta}}(\kappa) = -\frac{\sum_{i=1}^{\alpha} s_i}{\sum_{i=\beta}^{p} s_i},$$

which is constant within $R_{\alpha,\beta}$. In other words, the solution path is piecewise linear in the u-v plane.

Example 2. *For case 3, we have*

$$\frac{du_{\alpha,\beta}}{d\kappa} = \frac{\sum_{i=\beta}^{p} s_{p-i+1} - 2(p - \beta + 1)\kappa u}{\alpha + \kappa^2(p - \beta + 1)},$$

whose general solution is given by

$$u_{\alpha,\beta}(\kappa) = K \exp\left(\frac{\sum_{i=\beta}^{p} s_{p-i+1}}{\sqrt{\alpha(p - \beta + 1)}} \tan^{-1}(\kappa\sqrt{\alpha^{-1}(p - \beta + 1)}) + \log(\alpha + (p - \beta + 1)\kappa^2) \right), \quad (7)$$

for some constant $K > 0$.

Will the piecewise smooth solution path above be continuous as well? The concern is that at the boundary of the rectangle $R_{\alpha,\beta}$ where a small change of κ indeed alters α and/or β, there may be a jump in the path. The following lemma shows that this will not happen.

Lemma 1. *Suppose for some $\tilde{\kappa}$ with $(u_{\alpha,\beta}(\tilde{\kappa}), v_{\alpha,\beta}(\tilde{\kappa})) \in \mathbf{int} R_{\alpha,\beta}$. Let $\bar{\kappa} = \sup\{\kappa : (u_{\alpha,\beta}(\kappa), v_{\alpha,\beta}(\kappa)) \in R_{\alpha,\beta}\}$. Then the point $(u_{\alpha,\beta}(\bar{\kappa}), v_{\alpha,\beta}(\bar{\kappa}))$ coincides with either $(u_{\alpha-1,\beta}(\bar{\kappa}), v_{\alpha-1,\beta}(\bar{\kappa})) \in R_{\alpha-1,\beta}$, $(u_{\alpha,\beta+1}(\bar{\kappa}), v_{\alpha,\beta+1}(\bar{\kappa})) \in R_{\alpha,\beta+1}$, or $(u_{\alpha-1,\beta+1}(\bar{\kappa}), v_{\alpha-1,\beta+1}(\bar{\kappa})) \in R_{\alpha-1,\beta+1}$ exclusively.*

The proof is given in Appendix 1.

We have so far seen that the solution path is continuous and piecewise smooth, and how the curve pieces can be computed and traced. The remaining task is to determine the initial point

the path. The initial point can be obviously chosen to the point that corresponds to $\kappa = 1$, i.e., we need to find α and β such that $(u_{\alpha,\beta}(1), v_{\alpha,\beta}(1)) \in R_{\alpha,\beta}$. Note in this case that the closure of the desired $R_{\alpha,\beta}$ should intersect with the line $v = u$. By construction, this occurs if and only if $\alpha = \beta - 1$. Then, from (4) with $\kappa = 1$, it follows that

$$\sum_{i=1}^{p} l_i'(u) = \sum_{i=1}^{p} d_i = p\bar{d}, \quad \text{where} \quad \bar{d} = \frac{1}{p}\sum_{i=1}^{p} d_i, \tag{8}$$

and $u^\star(1)$ is found by solving this equation. In particular, if $l_i = l$ for $i = 1, \ldots, p$, then

$$u^\star(1) = (l')^{-1}(\bar{d}),$$

where $(l')^{-1}$ is the generalized inverse of l', which exists because l' is nondecreasing. Thus for case 1 we obtain $u^\star(1) = 1/\bar{s}$, and for case 3 we have $u^\star(1) = \bar{s}$.

Combining Lemma 1 and the above discussion, we are ready to fully describe the entire solution path, as stated in the following theorem.

Theorem 2. *If l_i, $i = 1, \ldots, p$, are closed convex and twice differentiable, the lower truncation value $u^\star(\kappa)$ for the optimal eigenvalue* (3) *for problem* (2)*, together with the upper truncation value $v(\kappa) = \kappa u(\kappa)$ traces a piecewise smooth path on the u-v plane as the regularization parameter κ varies. The resulting solution path is given by the solutions of the series of ordinary differential equations* (5)*, and its slope is discontinuous only when it intersects the vertical lines $u = \tilde{\lambda}_1, \ldots, \tilde{\lambda}_p$ or horizontal lines $v = \tilde{\lambda}_1, \ldots, \tilde{\lambda}_p$. The initial point of this path is found by solving* (8)*, corresponding to $\kappa = 1$. This initial point as well as the entire path can be found in $O(p)$ operations (Algorithm 1).*

Proof. Line 4 of Algorithm 1 takes $O(p)$ operations. In the loop, either of the conditions in Lines 11 and 12 must be met for each iteration. Thus for each value of $\alpha = 1, 2, \ldots, p$, at most one value of $\beta \in \{1, 2, \ldots, p\}$ is considered. This takes $O(p)$ time. □

Remark 3. *Algorithm 1 terminates if $v^\star = \tilde{\lambda}_{p-r+1}$, where*

$$\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_r > \tilde{\lambda}_{r+1} = \cdots = \tilde{\lambda}_p,$$

Algorithm 1 Solution path algorithm for orthogonal L

Set $\kappa_{\text{new}} \leftarrow 1$
Find $u^\star_{\text{new}} = v^\star_{\text{new}}$ by solving (8)
Find α such that $\tilde{\lambda}_{p-\alpha+1} < u^\star_{\text{new}} \leq \tilde{\lambda}_{p-\alpha}$; set $\beta \leftarrow \alpha + 1$
Set $\mathcal{K} \leftarrow \{\kappa_{\text{new}}\}$, $\mathcal{I} \leftarrow \{(\alpha, \beta)\}$
While ($\alpha \geq 1$ and $\beta \leq p$)
    Compute $u_{\alpha,\beta}(\kappa)$ by solving (5)
    Set $\kappa_u \leftarrow \inf\{\kappa \geq \kappa_{\text{new}} : u_{\alpha,\beta}(\kappa) = \tilde{\lambda}_{p-\alpha+1}\}$
    Set $\kappa_v \leftarrow \inf\{\kappa \geq \kappa_{\text{new}} : \kappa u_{\alpha,\beta}(\kappa) = \tilde{\lambda}_{p-\beta+1}\}$
    Set $\kappa_{\text{new}} \leftarrow \min(\kappa_u, \kappa_v)$
    $\mathcal{K} \leftarrow \mathcal{K} \cup \{\kappa_{\text{new}}\}$, $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\alpha, \beta)\}$
    If $u_{\alpha,\beta}(\kappa_{\text{new}}) = \tilde{\lambda}_{p-\alpha+1}$ then $\alpha \leftarrow \alpha - 1$
    If $\kappa_{\text{new}} u_{\alpha,\beta}(\kappa_{\text{new}}) = \tilde{\lambda}_{p-\beta+1}$ then $\beta \leftarrow \beta + 1$
Return $\mathcal{K}, \mathcal{I}$

understanding $\tilde{\lambda}_{p+1} = -\infty$. This includes the case when S is singular, i.e.,

$$s_1 \geq \cdots \geq s_r > 0 = s_{r+1} = \cdots = s_p.$$

For case 1, using the fact that the solution path is piecewise linear in the u-v plane, a simple geometric algorithm can be devised. This is shown in Algorithm 2.

Algorithm 2 Solution path algorithm for case 1

Set $\kappa_{\text{new}} \leftarrow 1$, $u^\star_{\text{new}} = v^\star_{\text{new}} = 1/\bar{s}$
Find α such that $l_\alpha > \bar{l} \geq l_{\alpha+1}$; set $\beta \leftarrow \alpha + 1$
Set $\mathcal{K} \leftarrow \{\kappa_{\text{new}}\}$, $\mathcal{U} \leftarrow \{u^\star\}$, $\mathcal{V} = \{v^\star\}$
While ($\alpha \geq 1$ and $\beta \leq p$)
    $t \leftarrow -(\sum_{i=1}^{\alpha} l_i)/(\sum_{i=\beta}^{p} l_i)$
    $\bar{R}_{\alpha,\beta} \leftarrow \{(u, v) : 1/l_\alpha \leq u \leq 1/l_{\alpha+1}$ and $1/l_{\beta-1} \leq v \leq 1/l_\beta\}$
    $(u^*, v^*) \leftarrow$ intersection between line passing $(u^\star_{\text{new}}, v^\star_{\text{new}})$ of slope t
                and boundary of $\bar{R}_{\alpha,\beta}$, with $u^* < u^\star_{\text{new}}$
    $\kappa_{\text{new}} \leftarrow v^*/u^*$, $u^\star_{\text{new}} \leftarrow u^*$, $v^\star_{\text{new}} \leftarrow v^*$
    $\mathcal{K} \leftarrow \mathcal{K} \cup \{\kappa_{\text{new}}\}$, $\mathcal{U} \leftarrow \mathcal{U} \cup \{u^\star_{\text{new}}\}$, $\mathcal{V} \leftarrow \mathcal{V} \cup \{v^\star_{\text{new}}\}$
    If $u^* = 1/s_\alpha$ then $\alpha \leftarrow \alpha - 1$
    If $v^* = 1/s_\beta$ then $\beta \leftarrow \beta + 1$
Return $\mathcal{K}, \mathcal{U}, \mathcal{V}$

3 Solution procedure for orthogonally variant $L(\Omega)$

With an additional sparsity-incuding penalty, problem (2) can be compactly written

$$\begin{array}{ll} \text{minimize} & h_1(\Omega) + h_2(\Omega) \\ \text{subject to} & \Omega \in \mathcal{C}_\kappa, \end{array} \tag{9}$$

where $h_1(\Omega) = L(\Omega) - \mathbf{Tr}(\Omega f(S)$ and $h_2(\Omega) = \mu|\Omega|_1 = \mu \sum_{i<j} |\Omega_{ij}|$. To be specific,

$$h_1(\Omega) = \begin{cases} -\log\det\Omega_D + (1/2)\mathbf{Tr}(\Omega S\Omega), & \text{case 4,} \\ (1/2)\mathbf{Tr}(\Omega S\Omega) - \mathbf{Tr}(\Omega), & \text{case 5.} \end{cases}$$

Problem (9) can be equivalently written

$$\text{minimize } h_1(\Omega) + h_2(\Omega) + \mathcal{I}_{\mathcal{C}_\kappa}(\Omega),$$

where

$$\mathcal{I}_{\mathcal{C}_\kappa}(\Omega) = \begin{cases} 0, & \Omega \in \mathcal{C}_\kappa \\ +\infty, & \text{otherwise.} \end{cases}$$

is the indictor function of the set $\mathcal{C}_\kappa$. Because both h_1 and h_2 are not orthogonally invariant, it is not obvious how to handle this spectral constraint set efficiently. The key idea here is to utilize the fact that the proximal operator of the indicator function $\mathcal{I}_{\mathcal{C}_\kappa}$, that is, the orthogonal projection to $\mathcal{C}_\kappa$, is efficiently computed using Algorithm 1. For $X \in \mathbb{S}^p$, where $\mathbb{S}^p$ is the space of $p \times p$ symmetric matrices, the proximal operator is defined as follows.

$$\mathcal{P}_{\mathcal{C}_\kappa}(X) = \underset{\tilde{X}\in\mathbb{S}^p}{\operatorname{argmin}} \mathcal{I}_{\mathcal{C}_\kappa}(\tilde{X}) + \frac{1}{2t}\|\tilde{X} - X\|_F^2, \quad t > 0. \tag{10}$$

The optimization problem involved in the right hand side of (10) is

$$\begin{array}{ll} \text{minimize} & (1/2)\|\tilde{X} - X\|_F^2 \\ \text{subject to} & \tilde{X} \in \mathcal{C}_\kappa, \end{array}$$

i.e., case 3. Thus, Algorithm 1 gives the entire solution to (10) for all $\kappa \geq 1$ in $O(p)$ operations, with the smooth pieces has a closed form given in (7), given the spectral decomposition of X.

Now (9) can be solved by using Dykstra's alternating projection algorithm (Lange, 2013, Ch. 15):

$$\begin{aligned}
\Omega^{(k+1/2)} &:= \operatorname*{argmin}_{\Omega \in \mathbb{S}^p} h_1(\Omega) + h_2(\Omega) + (1/2)\|\Omega - \bar{\Omega}^{(k)}\|_F^2 \\
\bar{\Omega}^{(k+1/2)} &:= 2\Omega^{(k+1/2)} - \bar{\Omega}^{(k)} \\
\Omega^{(k+1)} &:= \mathcal{P}_{\mathcal{C}_\kappa}(\bar{\Omega}^{(k+1/2)}) \\
\bar{\Omega}^{(k+1)} &:= \bar{\Omega}^{(k)} + \Omega^{(k+1)} - \Omega^{(k+1/2)},
\end{aligned} \tag{11}$$

which is an instance of the Douglas-Rachford operator splitting algorithm (Eckstein and Bertsekas, 1992); converges is guaranteed if $h_1(\Omega) + h_2(\Omega)$ is closed convex, which holds for cases 4 and 5.

For case 4, the subproblem (11) is to solve

$$\text{minimize} \quad -\log\det\Omega_D + (1/2)\mathbf{Tr}(\Omega(S + (1/2)I)\Omega) - \mathbf{Tr}(\Omega\bar{\Omega}^{(k)}) + \mu|\Omega|_1,$$

which is yet another CONCORD problem. This problem can be efficiently solved via the block coordinate descent (Khare et al., 2015), or proximal gradient methods (Oh, Dalal, Khare, and Rajaratnam, 2014).

For case 5, (11) reduces to a lasso program (Tibshirani, 1996):

$$\text{minimize} \quad (1/2)\mathbf{Tr}(\Omega(S + (1/2)I)\Omega) - \mathbf{Tr}(\Omega(I + \bar{\Omega}^{(k)})) + \mu|\Omega|_1,$$

which can again be efficiently solved via proximal gradient methods (Beck and Teboulle, 2009).

Illustration To illustrate the effect of the condition number regularization, we generated $n = 200$ samples from $p = 10$ dimensional multivariate normal distribution with zero mean and inverse covariance matrix Ω such that $\Omega_{ii} = 1$ for $i = 1, \ldots, p$ and $\Omega_{15} = \Omega_{51} = \Omega_{26} = \Omega_{62} = .99$. We compared the estimated Ω obtained using the CONCORD-ISTA algorithm (Oh et al., 2014) with

sparisty level $\mu = 0.1$ and that using the alternating projection algorithm of this section, where the upper bound for the condition number is set to 10 and the same CONCORD-ISTA is used for the subproblem (11). With the tolerance for the relative change of the estimates set as 1×10^{-6} (the meanings of the relative change are not the same between these two, though), the former terminated within 1000 iterations, and the latter within 503 iterations, where the inner CONCORD-ISTA is ran up to 100 iterations for each outer iteration. Both methods gave a similar sparsity pattern for the estimates (Figure 1). However, the inverse covariance matrix obtained using the CONCORD loss only is on the vicinity of singularity, with the minimum eigenvalue of 0.0102. The maximum eigenvalue was 1.98, giving the condition number of 194. On the other hand, the CONCORD loss combined with the condition number regularization yielded the minimum eigenvalue of 0.114, more than 10 times greater than the pseudo-likelihood-only counterpart, while the maximum eigenvalue was moderately reduced to 1.14. (Thus the condition number bound of 10.0 was retained.) The eigenvalue distributions of both cases are shown in Figure 2.

4 Conclusion

We have considered imposing a condition number constraint to regularize the estimator of the covariance of inverse covariance matrix of a population distribution under various loss criteria. For the losses that consists of an orthogonally invariant term and an inner product with a function of the sample covariance matrix, the problem reduces essentially that of the eigenvalues of the estimator, and the entire solution path with respect to the degree of condition number regularization can be obtained. If the involved ordinary differential equation admits a closed form solution, then the path can be obtained at the same cost as finding the estimator for a fixed regularization parameter. For other losses, an operator splitting scheme can be employed to find the estimator, hence the problem is scalable. At the core of this scheme lies the fact that the projection operator to the set of matrices with bounded conditio numbers allows path solutions, due to its orthogonal invariance.

The most expensive part in computing the solution paths is the spectral decomposition. As noted by Chi and Lange (2014), randomized algorithms such as random projection to lower dimensional subspaces may provide a computational relief (Mahoney, 2011). These approaches incurs

a small loss in accuracy, thus a possible research direction is to handle inexact solutions to the optimization subprolems in the alternating projection algorithm properly.

Appendix 1

Proof of Theorem 1. First note that both $L(\Omega)$ and $\mathcal{C}_\kappa$ are orthogonally invariant, hence depends only on the eigenvalues of Ω. Suppose the spectral decomposion of Ω is $U\Lambda U^T$, $U^TU = UU^T = I$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p$. For the trace part of the objective, the von Neumann-Fan inequality (Mirsky, 1975; Farrell, 1985; Lange, 2013, Appendix A.4) asserts that

$$\mathbf{Tr}(\Omega f(S)) \leq \mathbf{Tr}(\Lambda D) = \sum_{i=1}^{p} \lambda_i d_i,$$

with equality if and only if $V = U$. Thus problem (2) reduces to a $p+1$-variate problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{p} l_i(\lambda_i) - d_i\lambda_i \\ \text{subject to} & u \leq \lambda_i \leq \kappa u, \quad i = 1, \ldots, p, \\ & \lambda_1 \geq \cdots \geq \lambda_p, \end{array} \tag{12}$$

where the variables are $\lambda_1, \ldots, \lambda_p$ and u. The last order constraint can be removed, because of the following. Without the order constraint, for a fixed $u > 0$, the reduced problem (12) becomes separable in λ_i; it suffices to solve

$$\begin{array}{ll} \text{minimize} & l_i(\lambda_i) - d_i\lambda_i \\ \text{subject to} & u \leq \lambda_i \leq \kappa u \end{array} \tag{13}$$

for each $i = 1, \ldots, p$. Convexity of the objective in (13) ensures that the minimum is attained at

$$\lambda_i^*(u) = \max(u, \min(\tilde{\lambda}_i, \kappa u)),$$

where $\tilde{\lambda}_i = \operatorname{argmin}_\lambda l_i(\lambda) - d_i\lambda$. The optimality condition for $\tilde{\lambda}_i$ is given by

$$d_i \in \partial l_i(\tilde{\lambda}_i) \iff \tilde{\lambda}_i \in \partial g^*(d_i),$$

where $\partial f(x)$ denotes the subdifferential of f at x, and $g^*(v) = \sup\langle\lambda, u\rangle - g(\lambda)$, the convex conjuate of $g(\lambda)$. Monotoniciy of the subdifferential operator ensures that $\tilde{\lambda}_i$s perserve the order of d_is, i.e., $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_p$. It follows that $\lambda_1^*(u) \geq \cdots \geq \lambda_p^*(u)$, hence (12) reduces to a univariate minimization problem over u

$$\text{minimize} \sum_{i=1}^{p} l_i(\lambda_i^*(u)) - d_i\lambda_i^*(u). \tag{14}$$

The solution to (14), $u^\star$, must satisify

$$\lambda_{p-i+1}^*(u^\star) = \begin{cases} u^\star, & i = 1, \ldots, \alpha^\star, \\ \tilde{\lambda}_i, & i = \alpha^\star + 1, \ldots, \beta^\star - 1, \\ \kappa u^\star, & i = \beta^\star, \ldots, p, \end{cases}$$

where $\alpha^\star$ and $\beta^\star$ are such that $\tilde{\lambda}_{p-\alpha^\star+1} < u \leq \tilde{\lambda}_{p-\alpha^\star}$ and $\tilde{\lambda}_{p-\beta^\star+2} \leq \kappa u^\star < \tilde{\lambda}_{p-\beta^\star+1}$. To find $u^\star$, for $\alpha \in \{1, \ldots, p-1\}$ and $\beta \in \{2, \ldots, p\}$, define

$$\lambda_{p-i+1}^{\alpha,\beta}(u) = \begin{cases} u, & i = 1, \ldots, \alpha, \\ \tilde{\lambda}_i, & i = \alpha + 1, \ldots, \beta - 1, \\ \kappa u, & i = \beta, \ldots, p, \end{cases}$$

and

$$u_{\alpha,\beta} = \operatorname*{argmin}_u \sum_{i=1}^{p} l_{p-i+1}(\lambda_{p-i+1}^{\alpha,\beta}(u)) - d_{p-i+1}\lambda_{p-i+1}^{\alpha,\beta}(u) = \operatorname*{argmin}_u l_{\alpha,\beta}(u).$$

By construction, $u_{\alpha,\beta}$ coincides with $u^\star$ if and only if

$$\tilde{\lambda}_{p-\alpha+1} < u_{\alpha,\beta} \leq \tilde{\lambda}_{p-\alpha} \quad \text{and} \quad \tilde{\lambda}_{p-\beta+2} \leq \kappa u_{\alpha,\beta} < \tilde{\lambda}_{p-\beta+1}. \tag{15}$$

or $(u_{\alpha,\beta}, \kappa u_{\alpha,\beta}) \in R_{\alpha,\beta}$. Because $R_{\alpha,\beta}$s partition the u-v plane into $(p+2)^2$ regions and $(u_{\alpha,\beta}, \kappa u_{\alpha,\beta})$ is on the line $v = \kappa u$, an obvious algorithm to find the pair (α, β) that satisfies the condition (15) is to keep track of the rectangles $R_{\alpha,\beta}$ that intersect this line. To see that this algorithm takes $O(p)$ operations, start from the origin of the u-v plane, increase u and v along the line $v = \kappa u$. Since $\kappa \geq 1$, if the line intersects $R_{\alpha,\beta}$, then the next intersection occurs in one of the three rectangles: $R_{\alpha+1,\beta}$, $R_{\alpha,\beta+1}$, and $R_{\alpha+1,\beta+1}$. Therefore after finding the first intersection (which is on the line $u = \tilde{\lambda}_1$), the search requires at most $2p$ tests to satisfy condition (15). Finding the first intersection takes at most p tests. □

Proof of Lemma 1. Increase κ from $\bar{\kappa}$. Suppose the curve passing the point $(u_{\alpha^\star,\beta^\star}(\tilde{\kappa}), v_{\alpha^\star,\beta^\star}(\tilde{\kappa}))$ meets the left side (but not inclusive) $\{(u,v) : u = \tilde{\lambda}_{p-\alpha+1}\}$ of $R_{\alpha,\beta}$ before it meets the upper side (also not inclusive) $\{(u,v) : v = \tilde{\lambda}_{p-\beta+1}\}$. Then, taking the limit of both sides of (4) as $\kappa \nearrow \bar{\kappa}$, and by continuity of $u_{\alpha,\beta}(\kappa)$, we have

$$\sum_{i=1}^{\alpha} l'_{p-i+1}(\tilde{\lambda}_{p-\alpha+1}) + \bar{\kappa} \sum_{i=\beta}^{p} l'_{p-i+1}(\bar{\kappa}\tilde{\lambda}_{p-\alpha+1}) = \sum_{i=1}^{\alpha} d_{p-i+1} + \bar{\kappa} \sum_{i=\beta}^{p} d_{p-i+1}. \tag{16}$$

Optimality of $\tilde{\lambda}_{p-\alpha+1}$ (see (13)) and continuity of l'_{p-i+1} asserts that

$$l'_{p-i+1}(\tilde{\lambda}_{p-\alpha+1}) = d_{p-\alpha+1}.$$

Thus (16) is equivalent to

$$\sum_{i=1}^{\alpha-1} l'_{p-i+1}(\tilde{\lambda}_{p-\alpha+1}) + \bar{\kappa} \sum_{i=\beta}^{p} l'_{p-i+1}(\bar{\kappa}\tilde{\lambda}_{p-\alpha+1}) = \sum_{i=1}^{\alpha-1} d_{p-i+1} + \bar{\kappa} \sum_{i=\beta}^{p} d_{p-i+1}.$$

In other words,

$$\tilde{\lambda}_{p-\alpha+1} = u_{\alpha-1,\beta}(\bar{\kappa})$$

and $(u_{\alpha,\beta}(\bar{\kappa}), v_{\alpha,\beta}(\bar{\kappa})) = (u_{\alpha-1,\beta}(\bar{\kappa}), v_{\alpha-1,\beta}(\bar{\kappa})) \in R_{\alpha-1,\beta}$. If the curve meets the upper side before

the left side of $R_{\alpha,\beta}$, we have

$$\sum_{i=1}^{\alpha} l'_{p-i+1}(\tilde{\lambda}_{p-\beta+1}/\bar{\kappa}) + \bar{\kappa}\sum_{i=\beta}^{p} l'_{p-i+1}(\tilde{\lambda}_{p-\beta+1}) = \sum_{i=1}^{\alpha} d_{p-i+1} + \bar{\kappa}\sum_{i=\beta}^{p} d_{p-i+1},$$

$$l'_i(\tilde{\lambda}_{p-\beta+1}) = d_{p-\beta+1},$$

and thus

$$\sum_{i=1}^{\alpha} l'_{p-i+1}(\tilde{\lambda}_{p-\beta+1}/\bar{\kappa}) + \bar{\kappa}\sum_{i=\beta+1}^{p} l'_{p-i+1}(\tilde{\lambda}_{p-\beta+1}) = \sum_{i=1}^{\alpha} d_{p-i+1} + \bar{\kappa}\sum_{i=\beta+1}^{p} d_{p-i+1}$$

to have $(u_{\alpha,\beta}(\bar{\kappa}), v_{\alpha,\beta}(\bar{\kappa})) = (u_{\alpha,\beta+1}(\bar{\kappa}), v_{\alpha,\beta+1}(\bar{\kappa})) \in R_{\alpha,\beta+1}$. The final case, that the curve meets the upper left corner of $R_{\alpha,\beta}$, is the combination of previous two cases, and it follows that $(u_{\alpha,\beta}(\bar{\kappa}), v_{\alpha,\beta}(\bar{\kappa})) = (u_{\alpha-1,\beta+1}(\bar{\kappa}), v_{\alpha-1,\beta+1}(\bar{\kappa})) \in R_{\alpha-1,\beta+1}$. □

References

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences 2*(1), 183–202.

Chi, E. C. and K. Lange (2014). Stable estimation of a covariance matrix guided by nuclear norm penalties. *Computational Statistics & Data Analysis 80*, 117–128.

Eckstein, J. and D. P. Bertsekas (1992). On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming 55*(1-3), 293–318.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Farrell, R. H. (1985). *Multivariate calculation: Use of the continuous groups*. Springer.

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015). A convex pseudolikelihood framework for high

dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(4), 803–825.

Lange, K. (2013). *Optimization* (2 ed.). Springer.

Mahoney, M. (2011). Randomized algorithms for matrices and data. *Foundation and Trends in Machine Learning 3*(2), 123–224.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436–1462.

Mirsky, L. (1975). A trace inequality of john von neumann. *Monatshefte für Mathematik 79*(4), 303–306.

Oh, S., O. Dalal, K. Khare, and B. Rajaratnam (2014). Optimization methods for sparse pseudo-likelihood graphical model selection. In *Advances in Neural Information Processing Systems*, pp. 667–675.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*(486), 735–746.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 427–450.

Zhang, T. and H. Zou (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, ast059.

Zhao, P., G. Rocha, and B. Yu (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 3468–3497.

$$\begin{bmatrix}
1.00 & & & & 0.99 & & & & & \\
& 1.00 & & & & 0.99 & & & & \\
& & 1.00 & & & & & & & \\
& & & 1.00 & & & & & & \\
0.99 & & & & 1.00 & & & & & \\
& 0.99 & & & & 1.00 & & & & \\
& & & & & & 1.00 & & & \\
& & & & & & & 1.00 & & \\
& & & & & & & & 1.00 & \\
& & & & & & & & & 1.00
\end{bmatrix}$$

(a) True Ω

$$\begin{bmatrix}
1.03 & 0.00 & 0.00 & -0.00 & 1.01 & -0.01 & 0.01 & -0.00 & 0.01 & 0.00 \\
0.00 & 0.99 & 0.00 & -0.00 & 0.01 & 0.98 & 0.00 & 0.00 & 0.00 & -0.00 \\
0.00 & 0.00 & 1.04 & -0.01 & -0.00 & -0.00 & -0.06 & 0.00 & -0.01 & 0.00 \\
-0.00 & -0.00 & -0.01 & 0.93 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.01 & 0.01 & -0.00 & 0.00 & 1.02 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 \\
-0.01 & 0.98 & -0.00 & 0.00 & 0.00 & 0.99 & 0.00 & -0.01 & -0.03 & 0.00 \\
0.01 & 0.00 & -0.06 & 0.00 & 0.00 & 0.00 & 0.93 & 0.00 & 0.00 & 0.00 \\
-0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.01 & 0.00 & 0.93 & 0.02 & 0.00 \\
0.01 & 0.00 & -0.01 & 0.00 & 0.00 & -0.03 & 0.00 & 0.02 & 1.00 & 0.00 \\
0.00 & -0.00 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99
\end{bmatrix}$$

(b) CONCORD estimate

$$\begin{bmatrix}
0.63 & 0.00 & 0.00 & -0.00 & 0.51 & -0.00 & 0.00 & -0.00 & 0.01 & 0.00 \\
0.00 & 0.63 & 0.00 & -0.00 & 0.00 & 0.51 & 0.00 & 0.00 & 0.01 & -0.00 \\
0.00 & 0.00 & 1.04 & -0.01 & -0.00 & -0.00 & -0.06 & 0.00 & -0.01 & 0.00 \\
-0.00 & -0.00 & -0.01 & 0.93 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.51 & 0.00 & -0.00 & 0.00 & 0.63 & 0.00 & -0.00 & 0.00 & 0.00 & 0.01 \\
-0.00 & 0.51 & -0.00 & 0.00 & 0.00 & 0.63 & 0.00 & -0.00 & -0.02 & 0.00 \\
0.00 & 0.00 & -0.06 & 0.00 & -0.00 & 0.00 & 0.93 & 0.00 & 0.00 & 0.00 \\
-0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.00 & 0.00 & 0.93 & 0.02 & 0.00 \\
0.01 & 0.01 & -0.01 & 0.00 & 0.00 & -0.02 & 0.00 & 0.02 & 1.00 & 0.00 \\
0.00 & -0.00 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99
\end{bmatrix}$$

(c) CONCORD estimate with an upper bound on condition number

Figure 1: Illustration of the effect of the condition number regularization on the CONCORD pseudo-likelihood graphical model section.

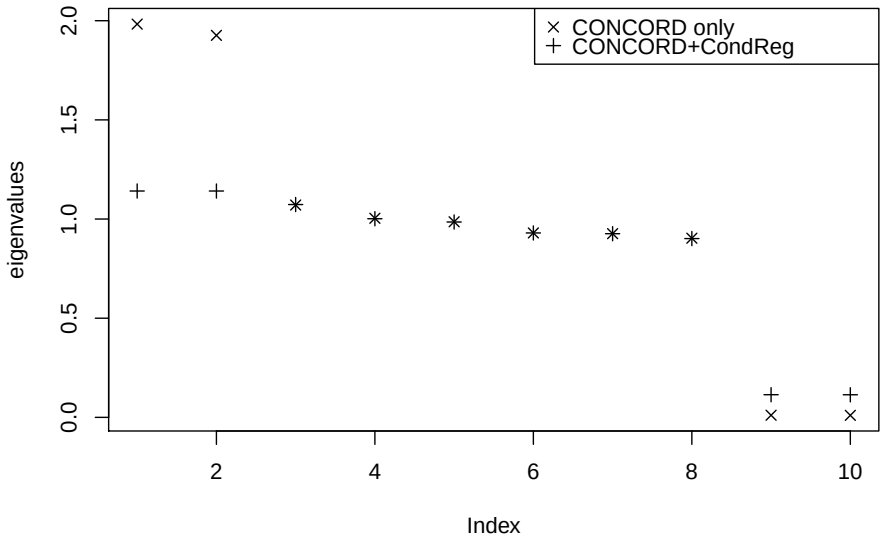

Figure 2: Distribution of the eigenvalues of the CONCOND-only inverse covariance matrix estimate (×), and CONCORD with condition number regularization (+).